DOI: 10.53469/jerp.2025.07(02).15

Construction and Pedagogical Validation of an AI-Empowered Multidimensional Assessment System for English Speech Course

Hongfang Xiao

Jiangxi University of Technology, Nanchang, Jiangxi, China
**Correspondence Author*

Abstract: ***This study focuses on the in-depth integration of artificial intelligence (AI) technology with English speech evaluation, with the objective of developing a multidimensional, quantifiable, and visualizable intelligent evaluation system, and validating its effectiveness through practical teaching applications. The research is framed within the context of multimodal learning theory, incorporating four core dimensions—language, content, expression, and interaction—and explores the innovative application of AI in data collection, model development, and result feedback. Through comparative experiments and empirical analysis, the paper demonstrates the practical value of the intelligent evaluation system in enhancing students' speech capabilities, stimulating learning motivation, and optimizing teaching efficiency. The findings offer both theoretical underpinnings and practical guidance for the reform of foreign language teaching evaluation in the age of artificial intelligence.***

Keywords: Artificial Intelligence, English Speech, Multidimensional Evaluation, Teaching Practice.

1. Introduction

In the context of accelerating globalization and increasingly frequent cross-cultural interactions, English speech proficiency has become a crucial measure of the overall competence of international talent. However, traditional English speech evaluation systems have long struggled with challenges such as excessive subjectivity, limited evaluative dimensions, and delayed feedback. Teachers often rely on subjective assessments based on experience, making it difficult to accurately quantify complex aspects such as pronunciation, content, delivery, and interaction. The limitations of manual grading also hinder students from receiving timely, targeted feedback, thereby constraining both teaching effectiveness and learning efficiency [1]. This challenge calls for a technological overhaul of educational evaluation systems, and the rapid advancement of artificial intelligence presents a promising new pathway.

AI technologies, with their powerful capabilities in data mining, multimodal analysis, and real-time feedback, are reshaping the landscape of educational evaluation. Technologies such as speech recognition, natural language processing, and computer vision enable precise capture and quantitative analysis of various speech features, including fluency, logical coherence, and body language. Meanwhile, emotional computing and machine learning algorithms offer deep insights into speakers' emotional expressions and their interaction with audiences, addressing the "invisible dimensions" often overlooked in traditional evaluations. By empowering evaluation systems, AI not only enhances objectivity and comprehensiveness but also facilitates the shift from a model where students passively receive evaluations to one in which they actively optimize their abilities, driven by instant feedback and personalized learning suggestions. This marks a new, dynamic phase for English speech instruction.

2. Theoretical Foundation: Integration of Multimodal Learning Theory and AI-Based Evaluation

2.1 Core Principles of Multimodal Learning Theory

Multimodal learning theory asserts that the learning process is not confined to the processing of information through a single sensory channel. Instead, it is a complex activity that involves the simultaneous engagement of multiple senses and media [2]. The theory emphasizes the collaborative role of different modalities—such as language, visual elements, and auditory cues—in transmitting information. In the context of English speech, this means that evaluation should extend beyond the quality of verbal expression to include non-verbal factors such as body language and eye contact. By integrating data from diverse modalities, learners are able to achieve a more holistic understanding and develop a deeper memory of the content. Consequently, multimodal learning theory provides essential theoretical support for creating a more comprehensive and scientifically grounded English speech evaluation system.

2.2 Advantages of AI-Based Evaluation

Artificial intelligence offers distinct advantages in data-driven analysis, multimodal evaluation, and real-time feedback. In the realm of English speech assessment, AI technologies—such as speech recognition, natural language processing, and computer vision—enable precise evaluation of various performance aspects, including linguistic ability, content organization, and emotional expression. Additionally, AI facilitates immediate feedback, allowing speakers to swiftly identify and address issues, thereby improving the quality of their presentations. This data-driven approach minimizes the influence of subjective bias and enhances the reliability and consistency of evaluation outcomes.

2.3 Integration of Multimodal Learning Theory with AI-Based Evaluation

2.3.1 Language Modality

In terms of the language modality, AI technologies can utilize speech recognition and natural language processing to assess various components of the speaker's performance, including pronunciation accuracy, intonation, vocabulary selection, grammatical structure, and fluency [3]. These technologies can detect subtle differences that traditional evaluation methods may overlook, providing more precise feedback through data analysis. For instance, AI systems are capable of identifying specific pronunciation errors or grammatical mistakes and offering targeted suggestions for improvement.

2.3.2 Visual Modality

The evaluation of the visual modality focuses on aspects such as body language, eye contact, emotional expression, and overall stage presence [4]. Through the application of computer vision and emotion recognition technologies, AI can analyze the speaker's facial expressions, gestures, and body posture to determine whether emotions are being effectively conveyed and if a strong connection with the audience is established. The integration of AI in this area allows for a more objective and quantifiable evaluation, offering speakers specific areas for improvement.

2.3.3 Text Modality

In the text modality, the evaluation centers on the quality of the speech content, considering factors such as relevance to the theme, logical coherence, persuasiveness, and originality. AI-driven text and sentiment analysis technologies are able to uncover hidden layers of meaning in the speech manuscript, assessing its ability to engage and persuade the audience. Furthermore, AI can analyze historical data to predict potential audience reactions, offering valuable insights that help speakers refine the content for greater impact.

2.3.4 Interaction Modality

The interaction modality evaluation focuses on assessing audience responses and the speaker's performance during the Q&A segment [5]. By utilizing AI-powered emotion recognition and speech analysis, it is possible to monitor audience sentiment in real time and record the speaker's effectiveness in responding to questions. This method not only reflects the actual impact of the speech but also facilitates more effective communication between the speaker and the audience.

3. Construction of a Multidimensional Evaluation System for English Speech Empowered by Artificial Intelligence

3.1 Design of Evaluation Dimensions

The introduction of artificial intelligence (AI) technology offers significant advancements in multidimensional, in-depth analysis for evaluating English speeches. Based on the theory of multimodal learning, the evaluation system is designed around four core dimensions: language, content, expression, and interaction, covering the essential elements of speech performance in a comprehensive manner.

In the language dimension, AI technologies, including speech recognition and natural language processing (NLP), allow for a precise evaluation of aspects such as pronunciation, intonation, vocabulary usage, grammatical accuracy, and fluency. Through speech recognition, the system can detect issues such as pronunciation errors, intonation imbalances, and fluctuations in speech rate, while NLP tools quantify vocabulary diversity and grammatical complexity [6]. This data-driven approach offers objective language assessments and can identify recurring errors, providing tailored guidance for further improvement.

The content dimension evaluation relies on text analysis and sentiment analysis techniques. These technologies facilitate an in-depth evaluation of the relevance, logical structure, persuasiveness, and originality of the speech content. For instance, logical analysis can examine the relationship between arguments and supporting evidence through semantic networks and keyword extraction [7]. Sentiment analysis evaluates the emotional tone embedded in the speech, providing insight into the content's ability to engage and affect the audience. Together, these technologies offer data-backed evaluations of the content's effectiveness and impact.

The expression dimension is concerned with non-verbal performance, including body language, eye contact, emotional expression, and stage presence. Computer vision plays a pivotal role in this dimension. By analyzing speech videos in real time, the system captures key aspects such as gesture frequency, facial expression variations, and gaze direction. When combined with sentiment analysis, these technologies assess the emotional accuracy and expressiveness of the speaker [8]. For example, the system can identify unconscious movements associated with nervousness and offer suggestions for improving stage presence and non-verbal communication.

The interaction dimension evaluates the engagement between the speaker and the audience. AI technologies, utilizing emotion recognition and speech recognition, analyze audience reactions and assess performance during Q&A sessions. The system captures facial expressions and body language from the audience to gauge their interest and engagement. In the Q&A segment, speech recognition records audience questions, while NLP techniques evaluate the accuracy and coherence of the speaker's responses. This multidimensional approach provides comprehensive feedback that helps the speaker improve their interactive communication skills.

3.2 Quantification of Evaluation Metrics

While the design of evaluation dimensions lays the foundation for the system, the quantification of evaluation metrics is essential for ensuring scientific rigor in the assessment. AI technologies, utilizing big data analytics and machine learning algorithms, convert evaluation criteria into quantifiable data models.

In the language dimension, the system extracts features such as pitch, speech rate, and pause frequency from speech recognition, while NLP calculates metrics such as vocabulary

diversity and grammatical error rates. These parameters undergo machine learning training and optimization, leading to an overall language proficiency score. In the content dimension, the system extracts keywords, thematic distributions, and sentiment from text, while logical structure analysis generates measurable indicators of relevance and coherence [9]. For the expression dimension, computer vision technologies extract non-verbal features such as facial expressions and gestures, which are then standardized to produce a comprehensive expression score. The interaction dimension is quantified by emotion recognition and speech recognition technologies, providing scores based on audience engagement and Q&A performance.

To ensure the accuracy and reliability of the evaluation model, the system is optimized through a combination of expert ratings and self-assessments. By comparing AI-generated scores with those of human experts, the system fine-tunes its parameters to improve the consistency and credibility of the evaluation results.

3.3 Visualization of Evaluation Results

The presentation of evaluation results plays a crucial role in their effectiveness. Leveraging AI technologies, the system employs data visualization techniques to transform complex evaluation data into clear, easily interpretable charts and reports, offering students actionable feedback and specific recommendations for improvement.

For the language dimension, the system generates visual representations such as waveform diagrams and speech rate variation curves, which clearly highlight pronunciation issues and variations in speech pace. In the content dimension, evaluation outcomes are represented through thematic distribution graphs and logical structure diagrams, helping students identify strengths and weaknesses in their speech content. For the expression dimension, visualization includes body movement heatmaps and emotional expression curves, offering students guidance on improving non-verbal communication [10]. The interaction dimension is visualized through audience emotional response graphs and radar charts that assess performance in the Q&A segment, enabling students to enhance their interaction with the audience.

The system also offers personalized learning suggestions based on the evaluation data. For example, in response to pronunciation issues in the language dimension, the system may recommend specific pronunciation practice exercises. For challenges in body language in the expression dimension, the system could suggest resources to improve stage presence. This personalized, data-driven feedback not only increases learning efficiency but also motivates students to take an active role in their own improvement.

4. Practical Validation of AI-empowered English Speech Instruction

4.1 Experimental Design

To validate the effectiveness and reliability of the AI-powered multidimensional evaluation system for English speeches, a semester-long instructional experiment was designed. The participants were two parallel classes of English majors at a university, with 30 students in each group. One group served as the experimental group, while the other acted as the control group. The experimental group utilized the AI-powered evaluation system for both teaching and assessment, while the control group relied on traditional methods of teacher grading and peer evaluations.

Throughout the experiment, both groups were required to complete four thematic speeches covering a variety of topics such as social issues, cultural differences, and technological innovation. In the experimental group, speech performance was analyzed using multidimensional data collection, including voice, text, video, and audience interaction data. In contrast, the control group's speeches were assessed by the teacher based on predefined grading criteria, supplemented by peer evaluations. To ensure the scientific integrity of the experiment, the initial English proficiency of both groups was assessed prior to the experiment. The results showed no significant differences in language skills or prior speaking experience between the two groups. After the experiment, data analysis was conducted to compare the improvement in speech abilities between the two groups, supported by surveys and interviews to gather student feedback on their experiences with the evaluation system.

4.2 Experimental Process

During the experiment, the experimental group made full use of the AI evaluation system's real-time feedback feature. After each speech, students could access a personalized evaluation report generated by the system, which provided insights into their performance across four dimensions: language, content, expression, and interaction. For instance, if a student's speech rate was too fast, causing unclear pronunciation, the system visually displayed the issue using speech waveform diagrams and rate fluctuation curves, offering targeted training recommendations to help manage speech pace. In the content dimension, the system identified logical gaps in the speech and suggested improvements based on thematic distribution and logical structure diagrams.

On the other hand, the control group followed traditional methods. The teacher evaluated the speeches based on a predefined scoring rubric and provided collective feedback during class. Peer evaluations were conducted, with students using a rating sheet provided by the teacher. While this method offered some feedback, the limited number of evaluation dimensions and delayed feedback made it difficult for students to receive immediate and specific suggestions for improvement.

Throughout the experiment, the research team recorded the speeches of both groups, including video, audio, text, and audience interaction data, which laid a strong foundation for subsequent data analysis.

4.3 Analysis of Experimental Results

After the experiment, the post-test data revealed that the experimental group outperformed the control group across all four evaluation dimensions—language, content, expression, and interaction. In the language dimension, students in the

experimental group showed significant improvement in pronunciation accuracy and speech rate control, with speech waveform diagrams indicating more consistent pacing. In the content dimension, the experimental group demonstrated enhanced logical coherence and persuasiveness in their speeches. The thematic distribution graphs showed a stronger connection between arguments and supporting evidence.

Analysis of the expression dimension showed notable improvements in body language and emotional expression in the experimental group. Data captured by computer vision technology indicated that students in this group used gestures more naturally, and their facial expressions were more aligned with the emotional tone of their speeches. In the interaction dimension, the experimental group displayed greater confidence in the Q&A sessions, with audience emotion recognition data indicating an increase in audience engagement and speech appeal.

Survey and interview results further confirmed the effectiveness of the AI evaluation system. Students in the experimental group commonly reported that the system's real-time feedback and personalized recommendations helped them quickly identify issues and find ways to improve. One student said, “The speech waveform diagram helped me realize that I was unconsciously speeding up my speech. After focused training, my pronunciation has become much clearer.” Another student mentioned, “The system’s logical structure diagram made me realize that my arguments were not sufficiently highlighted. After making adjustments, my speech became much more persuasive.”

In contrast, while the control group students did show some improvement, the lack of multidimensional, real-time feedback meant their progress was slower and less effective. One student from the control group commented, “The teacher’s feedback was helpful, but after each speech, I only received an overall score, making it difficult to pinpoint which areas needed improvement.”

4.4 Experimental Conclusion

The results of the practical validation demonstrate that the AI-powered multidimensional evaluation system for English speeches offers significant advantages in improving students' speech abilities and optimizing instructional outcomes. The system’s data-driven approach, coupled with real-time feedback, not only helps students rapidly identify areas for improvement but also enhances their motivation and ability to engage in autonomous learning.

Throughout the experiment, the AI evaluation system showed its comprehensive analysis capabilities in language, content, expression, and interaction, providing a scientific and efficient tool for English speech instruction. With its data-driven evaluation model, teachers can gain a more precise understanding of students’ learning needs, optimize lesson planning, and foster innovation in teaching methods.

This experiment provides empirical support for the application of AI in foreign language education and sets the stage for future research. Future studies could focus on refining the accuracy and reliability of the evaluation model, expanding the use of AI evaluation systems to larger-scale teaching practices, and exploring its potential in evaluating other language skills. These questions offer rich avenues for further investigation in future research.

5. Conclusion

This study presents an AI-powered multidimensional evaluation system for English speech, marking a transformative shift in educational assessment paradigms driven by technology. Traditional methods of evaluating English speech have long been limited by issues such as subjectivity, delayed feedback, and narrow focus. By integrating multimodal learning theory with AI technology, this research develops an intelligent evaluation system that spans four core dimensions—language, content, expression, and interaction—striking a balance between scientific rigor and practical applicability. The experimental data indicate that this system effectively captures the multifaceted performance of speakers, providing real-time, visual feedback that significantly enhances students’ speaking abilities and learning efficiency.

The core contributions of this study can be understood from three perspectives. Theoretically, the deep integration of multimodal learning theory and AI technology broadens the theoretical boundaries of educational assessment and introduces a new model for foreign language teaching evaluation. Methodologically, the multidimensional quantitative model and data visualization tools overcome the limitations of traditional assessment approaches, transitioning from subjective judgments to data-driven evaluations. Practically, the results of the teaching experiment validate the system’s dual role in promoting personalized learning and optimizing teaching strategies, offering a replicable model for the integration of AI in educational practice.

However, this study acknowledges certain limitations. The dataset used to train the evaluation model needs to be expanded to increase its applicability across broader contexts. Additionally, the stability of audience emotion recognition technology in complex real-world environments requires further refinement. Future research can explore several promising directions: first, investigating the potential of deep learning and generative AI to enhance the evaluation model’s ability to interpret complex linguistic features and non-verbal behaviors; second, fostering interdisciplinary collaboration between educational theory, psychology, and technology development to create a more human-centered evaluation framework; and third, expanding the system’s applications to broader domains such as international speech competitions and professional training to assess its societal impact and industrial relevance.

The profound integration of AI with education is an inevitable and irreversible trend. This study not only provides a sophisticated tool for English speech instruction but also urges educators to embrace technological transformation with an open and adaptive mindset. In this age of "human-machine collaboration," it is crucial to redefine the role of teachers—from knowledge transmitters to facilitators of learning and emotional empowerment. It is through this balanced approach, where technology and humanity intersect,

that we can truly realize the educational vision of “learning through assessment, teaching through assessment,” and nurture a new generation of communicators who are well-prepared for the digital era.

Acknowledgements

2024 Ministry of Education Industry-University Collaborative Education Project "Exploration and Practice of Multiple Teaching Modes of English Speech Empowered by Artificial Intelligence" (Project No.: 231100273142521)

References

[1] Sarsenbaeva Z, Uteshova Z.Principles of Teaching Karakalpak Students English Speech Etiquette [J]. Humanising Language Teaching, 2022, 24(4).
[2] Zheng J. On the application of intelligent speech aids in English teaching under multimedia environment[J].J. Comput. Methods Sci. Eng. 2021, 21:1999-2008.
[3] Li L, Dan L. On Teachers' "SOFT" Roles and the "WARE" Teaching Mode in College English Speech Course[J].Educational Research on Foreign Languages & Arts, 2010.
[4] Hashim M A, Mukhopadhyay S, Sahu J N, et al. Remediation technologies for heavy metal contaminated groundwater [J]. Journal of Environmental Management, 2011, 92(10):2355-2388.
[5] Fang G. Language-driven or Content-driven - on CBI Based Teaching Model of Public English Speech[J]. 2021.
[6] Cui W. Research on Speech Recognition and Feedback Technology in AI-Driven English Speaking Practice Platforms[C]//International Conference on Artificial Intelligence for Society. Springer, Cham, 2024.
[7] Yao X. Research on the Innovation of University English Teaching Mode Driven by Artificial Intelligence [J]. Applied Mathematics and Nonlinear Sciences, 2024, 9(1). DOI:10.2478/amns-2024-2917.
[8] Zhang X, Qin Q. Development and Application of College English Translation Teaching Software Based on Artificial Intelligence[C]//EAI International Conference, BigIoT-EDU. Springer, Cham, 2024. DOI:10.1007/978-3-031-63136-8_44.
[9] Kazu B Y, Kuvvetli M. The Influence of Pronunciation Education via Artificial Intelligence Technology on Vocabulary Acquisition in Learning English [J]. International Journal of Psychology and Educational Studies, 2023.
[10] Haoxin Y Q. Artificial intelligence speech recognition model for correcting spoken English teaching [J]. Journal of intelligent & fuzzy systems: Applications in Engineering and Technology, 2021, 40(2).